The Use of Naturalistic Reading Corpora for the Study of Pronoun and Coreference Resolution

Olga Seminck

Author: Olga Seminck
ORCID: 0000-0003-4617-5992

Title:
The Use of Naturalistic Reading Corpora for the Study of Pronoun and Coreference Resolution

Key Words:
pronoun resolution, coreference resolution, eye-tracking, naturalistic text reading, cognitive computational linguistics, psycholinguistics, processing cost

Abstract:
Naturalistic reading corpora are collections of texts that were not designed to be used in specific linguistic studies and that were read by participants whose eye-movements or reading time was measured. These resources are used to study the cognitive processing of linguistic phenomena naturally present in texts and they encourage the development of robust models of cognitive linguistic processing. These properties make the use of natural text corpora interesting for the study of pronoun and coreference resolution. In the psycholinguistic literature, many linguistic factors that have an influence on pronoun and coreference resolution have been identified but there is still a lot unknown about the interaction of these factors in naturalistic data. In addition, items used in psycholinguistic studies are short; therefore, naturalistic reading corpora are a resource to study pronoun and coreference resolution in realistic discourse. In this survey, we discuss the models for pronoun and coreference resolution that have been developed so far. We explain the methodological challenges related to the use of naturalistic data and speculate how such data can be used to evaluate theories of pronoun and coreference resolution and so lead to the development of broad coverage models in which various linguistic levels (syntax, semantics and discourse) are integrated.

Main Text:

§1 Introduction

In this survey, we discuss the study of pronoun and coreference resolution on naturalistic reading corpora. As this subject involves many disciplines and topics, we will first, in this introduction, explain what they all entail. Second, we will review in detail the models that have been developed so far and discuss what we can learn from their results. Lastly, we finish this article with a discussion about how cognitive computational models of pronoun and coreference resolution on naturalistic data contribute to theories about pronoun and coreference resolution in particular, and also about cognitive processing cost induced by linguistic input in general.

§1.1 Pronoun and Coreference Resolution

Pronoun resolution is the process of finding the antecedent of an anaphoric pronoun. Anaphoric means that the semantic interpretation of the pronoun is dependent on its antecedent (Van Deemter & Kibble, 2000). A simple example is given in 1. below.

1. *Ernest*$_{(antecedent)}$ *went into the kitchen and* ***he***$_{(anaphoric\ pronoun)}$ *found some snacks.*

Coreference resolution is a broader phenomenon that has a lot in common with pronoun resolution: it is the process of finding two linguistic expressions that refer to the same entity (Van Deemter & Kibble, 2000). However, whereas in an anaphoric relationship there is the criterion of dependence of the anaphor on its antecedent, in the case of coreference dependence is not required: in a text where the name '*Micky Mouse'* is mentioned multiple times, these mentions are coreferent, but not in an anaphoric relationship.

Pronoun resolution is the most frequently studied type of coreference resolution in the domain of psycholinguistics. Most studies on human pronoun resolution have used controlled experimental items: sentences that have been created for the experiment and that vary as little as possible from one another per experimental condition. This has led to a large literature on linguistic factors that have an influence on pronoun resolution, across various languages. For example, in English, subject antecedents are processed quicker than object antecedents (Crawley *et al*., 1990). Some other examples of much discussed factors are: the *first mention bias* (people tend to resolve an ambiguous pronoun to the antecedent that was mentioned first) (Gernsbacher & Hargreaves, 1988; Gernsbacher *et al*., 1989; Gernsbacher, 1990; Järvikivi *et al*., 2005); the *parallel function bias* (a preference to resolve an ambiguous pronoun to an antecedent that has the same syntactic function) (Maratsos, 1973; Sheldon, 1974; Smyth, 1994); and the preference for linking a pronoun to the closest antecedent (Clark & Sengul, 1979; Ehrlich & Rayner, 1983).

These factors have often been explained in the light of *salience*, making reference to *Centering Theory* (Grosz *et al*., 1983) and *Accessibility Theory* (Ariel, 1988). According to these theories, pronouns refer to discourse referents that are salient (which means easily retrievable from memory): they have a more prominent grammatical function, such as the subject; they are more recent; and they have a higher frequency in the text.

However, even though the factors that are found in classical psycholinguistic research can often be explained in the light of saliency theories, a gap remains between the very broad aim of these theories --- that explain saliency on a discourse level --- and the psycholinguistic experiments that test one linguistic factor at a time in an artificial, strongly reduced discourse composed of one, two or three sentences.

§1.2 Naturalistic Reading Corpora

There is thus a need for testing theories of pronoun and coreference resolution in natural discourse. Naturalistic reading corpora could help to achieve this goal. In naturalistic reading corpora, reading data (eye-tracking data or self-paced reading data) has been collected from people who read texts that are not experimental stimuli, for example, novels or newspaper articles. These texts, together with the reading data, form a naturalistic reading corpus and

enable the possibility of studying the cognitive load of pronoun and coreference resolution in a rich and non-reduced discourse.

The manner in which reading data is exploited is based on the following assumption: reading is an *online* process, meaning that the linguistic input is processed by the brain as soon as it is encountered, first as visual input, and then soon transformed into linguistic input consisting of words, sentences and finally a discourse. When a text is more challenging, for example, because of the use of long sentences, the processing thereof takes longer, hence the reading process is slower.

To understand what slower reading actually means, we have to take a closer look at the reading process: how does it work physically? While reading, the eyes do not scan the text in a fluid movement from left to right but make fixations that take a fraction of a second (for example, 150 ms) before jumping to the next fixation, a process called *saccades* (Rayner, 1998). The text only enters the brain during fixations, as vision is suppressed during the saccades.

When reading data is used to see which linguistic structures are processed with more ease or more difficulty, the pattern of fixations can be exploited in various ways (Rayner, 1998). One is to look at the fixation duration: longer fixation duration indicates slower reading. Fixation duration can be used to calculate different measurements of reading time, for example, the fixation durations of all fixations on a word can be summed to obtain the measure of *total reading time*. Another example of a measure is the *first pass reading time.* This is the sum of the durations of the fixations that fell between the first moment the word was fixated on until the gaze fell on another word. A second way in which to exploit reading data is to give special consideration to the times that readers don't continue reading new text, but actually look back, returning to earlier parts of the text. Eye-movements to the left are called *regressions*[1]*.* Regressions can be measured by reading metrics such as the *regression path reading time* and *second pass reading time*. A third example of a way to use reading data is by having a look at the density of fixations. Indeed, when the text is challenging, the saccadic length tends to decrease: the eyes make fixations that are closer to each other. On the other hand, in non-challenging texts many words are actually not fixated, the rate of non-fixated tokens drops in difficult texts (Brysbaert & Vitu, 1998). Therefore, a denser fixation pattern can also be associated with more cognitive difficulty.

§1.3 Computational Models of Cognitive Processing Cost

In studies about cognitive computational modelling on naturalistic reading corpora, a computer model tries to predict a reading metric by looking at the linguistic features of a text. For example, a model could predict the first pass reading time of words by measuring their frequency in a corpus. Computational models of cognitive processing cost exist for types of measurements other than reading metrics, for example: event related potentials recorded by EEG, or the BOLD-response measured during an fMRI experiment (Armeni *et al.,* 2017).

[1] If we talk about text that is written from left to right.

Cognitive processing cost models can be used to determine whether linguistic theories can make accurate predictions about the processing load of linguistic input. It is important to note that they can only evaluate a theory's plausibility by answering the following question: *given a linguistic input and the implementation of a theory, can the behaviour of humans be accurately predicted?* It should be clear that the answer depends heavily on the given implementation; when the result is negative, the implementation can be blamed and it cannot be considered as proof of the theory being wrong.

If we consider the different levels of cognitive modelling of Marr (1982) --- who defined three levels, going from a high level of abstract theoretical specifications to a low level of specifications about the physical realisations of the algorithm --- cognitive processing cost models are at the highest level, the computational level. They simulate the produced output by specifying the model's parameters but the algorithms that are used often don't claim to bear similarities with cognitive processes. For example, if a cognitive computational model uses, at some point, a context-free grammar parser (CFG-parser) to predict reading time, it does not have to claim that humans parse syntax in CFG-style.

Except when evaluating the plausibility of individual theoretical frameworks, cognitive computational modelling can be used to compare two competing theories. This can be illustrated by Demberg and Keller's 2008 study. They evaluated whether *Dependency Locality Theory* (Gibson, 2000) and *Surprisal Theory* (Hale, 2001; Levy, 2008) could predict reading times in the Dundee Corpus, a corpus of about 50K tokens from the newspaper *The Independent* for which eye-movements of ten native speakers of English were recorded (Kennedy *et al*., 2003).

According to Dependency Locality Theory, an integration cost can be calculated for discourse referents (which are basically nouns and verbs). The integration cost of a discourse referent depends on two factors. On the one hand, it is assessed whether the discourse referent is discourse new or discourse old; the former resulting in more processing cost than the latter. On the other hand, the integration cost of the discourse referent is determined by the number of intervening discourse referents between itself and its syntactic head: more referents lead to higher processing cost. No integration cost is predicted for words other than discourse referents.

According to Surprisal Theory, processing cost is related to how expected a given linguistic input is: more expected input results in less processing cost. Sentences are processed word by word, and for every new word its probability can be evaluated given the preceding context. To estimate a word's probability, language models can be used. Often, models that are sensitive to syntactic structure have been shown effective in predicting why some linguistic structures are more difficult to process than others (for example, *garden paths* and object relative clauses (Hale (2001)).

Demberg and Keller (2008) found that syntactic surprisal was a significant predictor of different reading times (first fixation duration, first pass, and total reading time). Dependency Locality Theory was not a significant predictor of reading times when applied to all words of the Dundee Corpus, but it was a significant predictor when only nouns and verbs

were considered. It seems that Dependency Locality does matter for discourse referents (nouns and verbs), but the integration cost of other syntactic categories cannot be ignored. This result shows that cognitive computational modelling on naturalistic corpora can be used as a means to evaluate a theory's *robustness*: its capacity to make relevant predictions taking into account all types of linguistic structures. A second interesting result was that surprisal and integration cost were not strongly correlated, suggesting that both these theories can be complementary.

In addition to the study of Demberg and Keller (2008), that we used to illustrate the methodology of cognitive computational modelling on naturalistic data, there are many other interesting studies (*e.g.* Lau *et al*., 2017; Brennan *et al*., 2016; Willems *et al*., 2015; Frank & Bod, 2011; Mitchell *et al*., 2010), which we cannot detail further here because of space restrictions. However, it is important to note that most cognitive computational models of language encode lexical or syntactic features, but models specifically predicting the processing cost of semantics or discourse phenomena are very scarce. Therefore, we believe that it is interesting to review the models proposed for pronoun and coreference resolution and to reflect on how, eventually, models can be developed that take into account various linguistic levels simultaneously.

§2 Survey of models of pronoun and coreference resolution tested on naturalistic reading corpora

There are a lot of interesting studies that have mostly been done on the topic of pronoun resolution and some on coreference resolution using reading data (*e.g.* Koornneef, 2008; Frank *et al*., 2007; Van Gompel & Majid, 2004; Sturt, 2003; Ehrlich & Rayner, 1983). But there are only --- as far as we know --- three computational models of pronoun and coreference resolution that have been evaluated on naturalistic reading data. As we believe that the naturalistic aspect and the evaluation on corpus is crucial, we will discuss only these studies in more depth.

§2.1 A Surprisal Metric of Syntax and Coreference

Dubey *et al*. (2013) intended to enrich syntactic surprisal with information about discourse referents. They present a broad coverage model --- called *the paired model* --- that integrates information about syntax and discourse referents. Syntactic surprisal is adapted according to whether a discourse referent is discourse old or discourse new. The authors posit that discourse new referents will induce more cognitive load for readers than discourse old referents, because discourse old referents are more frequent than ($P = 0.58$) than discourse new referents ($P = 0.41$).

The model uses two modules. The first is a part-of-speech tagger, based on a hidden Markov model (HMM) that is used to identify noun-phrases (that are all considered as discourse referents) and used to calculate surprisal. The second module is a coreference module that keeps in a cache a list of encountered mentions and determines, with simple rules, whether new mentions are coreferent with them. The probabilities outputted by the HMM are

adapted according to whether the coreference module estimated whether a discourse referent is discourse old or discourse new.

Dubey *et al*.'s (2013) model is evaluated on the Dundee Corpus. A *mixed effects model* was built that tries to predict the total reading time of every token in the corpus. The random factors of the model were random intercepts for participants and tokens in the corpus. Because of the use of naturalistic data, confound variables --- such as word frequency and word length --- must be statistically controlled (Armeni *et al*., 2017). Therefore, except for surprisal scores, the independent variables of this model were: the log-frequency of words, the number of characters in the word, the position of the word on the line, and the position of the line in the document.

To estimate the contribution of both syntactic surprisal and surprisal based on syntax *and* coreference information, different versions of the model were compared in a step-wise fashion: first, the fit of a model without any surprisal was compared to a model containing a predictor of syntactic surprisal only, and then this second model was compared to a model containing both syntactic and coreferential surprisal. Both the adding of syntactic and coreferential surprisal gave the model a better fit to the data, according to model comparison metrics such as the Aikake Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Also, the log-likelihood significantly improved as surprisal measures were added. In the final model, that contained both syntax and coreference surprisal, all predictor variables were statistically significant. This shows that, in addition to syntactic structure, the cognitive processing of a text is easier for discourse old referents.

§2.2 The Impact of Focus on Coreference Resolution

A recent study has examined the influence of focus effects on coreference resolution (Jaffe *et al*., 2018). In this study, the process of focussing is defined in the following way: "*Linguistic focus directs subjects' attention toward particularly salient or important discourse referents during sentence processing."* Some syntactic structures are known to let a focus effect emerge, for example, the it-cleft: the clefted clause becomes more salient, as in:

2. *It was malaria$_{focus}$ that Sophia got infected with in Tanzania.*

According to Jaffe *et al*. (2018), focussing plays an important role in coreference resolution: mentions of referents that are focussed induce less cognitive load and can therefore be processed quicker. To test their hypothesis, they investigated whether two metrics of focussing can predict reading times on corpus.

The first metric they propose is frequency-based. The underlying assumption is that discourse referents that are mentioned more frequently are more salient. The frequency-based metric is very simple: every discourse mention is scored by the number of times the discourse entity it refers to was previously mentioned in the text: a mention that is not coreferent with a previously mentioned entity gets the value of 0, a mention that refers to an entity that has been mentioned once before gets the value of 1, and a mention that refers to an entity mentioned twice before gets the value of 2, etc.

The second metric is based on recency. The assumption that underlies the mention-recency variable is that more recently mentioned entities are more salient and can be retrieved more easily from memory. Two versions of this metric were tested: the number of words between the mention and its referent (this variable is set to 0 for first mentions) and the number of intervening mentions.

The two focus-based metrics were tested on the reading times of the Natural Story Corpus (Futrell *et al.*, 2018). This corpus contains stories that sound natural, but that have been manipulated to contain low frequency phenomena of language. This enables linguists to study low-frequency phenomena in a naturalistic setting. The corpus comes with reading time data from 181 participants that performed a self-paced reading task thereon. Jaffe *et al*. (2018) annotated this resource with coreference chains.

The two focus-based metrics were tested in a mixed effects model that included the variable of frequency and the two variables of recency. Besides the factors based on the focus metrics, control factors were included in the model: word length, syntactic surprisal (using probability estimations from a Probabilistic Context Free Grammar parser), n-gram surprisal, and a variable called 'story position' that indicates what percentage of the story has already been read by the participant. Random slopes representing the participants were included for all factors except for syntactic surprisal because of convergence problems.

The model was first optimised on one-third of the corpus that was held out as a development set. There, it turned out that all the factors had significant contributions to the model's estimation except for the recency-based focus metrics. Therefore, these factors were left out in the final model that was tested on the remaining two-thirds of the corpus. The final model showed a highly significant but small effect of mention count: when a mention refers to an entity that has been mentioned more often, reading times are lower. This result is in line with their focussing hypothesis.

§2.3 A Cost Metric of Pronoun Resolution Based on Entropy

Seminck (2018) designed a cost metric to simulate the cognitive cost of pronoun resolution. The cost metric was based on the following hypothesis: more competition amongst the antecedent candidates of a pronoun leads to a higher processing load for humans. Competition is measured by the entropy over a probability distribution of all potential antecedent candidates.

Entropy is a measure from Information Theory (Shannon, 1948) that captures how much uncertainty there is in a given random variable (Thomas & Cover, 2006). A uniform probability distribution --- which reflects maximal uncertainty --- leads to the highest entropy. On the contrary, a distribution in which one antecedent is very likely and the others are not, leads to low entropy.

$$H(X) = -\sum_{i} p(i) \cdot \log_2\big(p(i)\big)$$

To measure the entropy for a pronoun, the pronoun resolution process is modelled as a random variable: all the potential antecedents ($a \in A$) are possible outcomes to which probability can be assigned. The cognitive cost (C) is proportional to the entropy over this random variable.

$$C(pro) \approx -\sum_{a \in A} p(pro = a) \cdot \log_2\big(p(pro = a)\big)$$

However, entropy increases by the number of outcome events in the random variable. In a setting of naturalistic data, the number of possible antecedents is expected to become higher and higher throughout the text. Therefore, the metric was slightly adapted to prevent giving systematically higher scores to the pronouns at the end of the texts. The adaptation of the metric consisted of taking the relative entropy between the entropy measured as proposed above (P), and the maximal entropy (Q): entropy that is obtained on a flat probability distribution that reflects maximal uncertainty. Relative entropy is often interpreted as the distance between probability distributions, even though relative entropy is not symmetrical: $H_{relative}(P||Q) \neq H_{relative}(Q||P)$ (Thomas & Cover, 2006). Because the entropy of pronoun resolution is compared to maximal entropy, the prediction is that less relative entropy leads to higher processing cost.

$$H_{relative}(P||Q) = \sum_{i \in P \wedge i \in Q} P(i) \log_2 \frac{P(i)}{Q(i)}$$

The challenge of implementing the cost metric is to determine what exactly an antecedent candidate is and how probability over antecedent candidates should be calculated. Note that this probabilistic view of pronoun resolution hypothesises that people attribute some probability to non-antecedents, even if the resolution of a given pronoun seems straightforward (non-ambiguous).

Seminck (2018) chose to use a state of the art probabilistic coreference resolution system (Lee *et al*., 2017) to find the set of antecedent candidates and to estimate their probabilities. Lee *et al*.'s 2017 system is an *end-to-end system*, which means that instead of taking the output from an NLP-pipeline architecture that performs tokenisation, parsing, etc., the system takes raw text and immediately performs coreference resolution.

Seminck (2018) tested whether the entropy metric was able to predict how humans read pronouns in the Dundee Corpus. All anaphorical pronouns[2] in the Dundee Corpus were previously identified (Seminck & Amsili, 2018), leading to a set of 1,109 items. An example of a pronoun-antecedent pair from the corpus can be found below in 3:

3. *All parties welcomed [the Woolf report]$_{(antecedent)}$ of 1991 yet none has acted on* ***it***$_{(anaphoric\ pronoun)}$ *to improve the education service in prison.*

[2] Other types of pronouns are, for example, non-referential pronouns (or pleonastic pronouns) and deictic pronouns.

A challenge that Seminck (2018) discusses is that pronouns are often skipped by participants, only about 40% is fixated in the Dundee Corpus (Barrett and Søgaard 2015; Seminck 2018). Therefore, she decided to study the skipping behaviour, because this measure leads to the least data sparsity, in contrast to reading times.

For every anaphorical pronoun relative entropy was calculated. The relative entropy was entered as an independent variable in a mixed effects model. This model also contained the following control factors: the length of the pronoun in characters, the word's frequency estimated on the British National Corpus, whether there was a comma attached to the pronoun, and whether there was a punctuation mark associated with the end of a sentence (full stop, exclamation mark, or question mark). Participants were entered as random intercepts and also as random slopes on the relative entropy metric variable. Items were modelled with random intercepts. All the independent variables were scaled. The model was estimated with the *brms* package (Bürkner 2017) in R (R Development Core Team, 2008) that implements mixed effect modelling with Bayesian statistics[3] (Downey, 2013).

The model showed that a higher relative entropy leads to less fixations on the pronoun (95% confidence interval between -0.13 and -0.01), a result in line with the hypothesis. All other factors also had their 95% confidence intervals not crossing 0, except for the factor that checked whether there was an end of sentence punctuation marker attached to the pronoun.

Seminck (2018) concludes that competition amongst antecedent candidates is a factor of influence on pronoun resolution and that the probabilistic view of resolution is plausible. Therefore, the presented cost metric is in line with other cost metrics based on Information Theory, such as surprisal.

§3 Discussion

When we compare the studies on pronoun and coreference resolution on naturalistic reading corpora, we can first observe that what both the studies of Dubey *et al.* (2013) and Jaffe *et al*. (2018) have in common is this: the fact that a discourse referent has been mentioned before, has an effect on its cognitive load. Dubey *et al*. (2013) noticed that lowering their surprisal metric for discourse old mentions improved the fit of their model; and Jaffe *et al*. (2018) found the number of times an entity was previously mentioned was a highly significant factor in their model for reading times in the Natural Story Corpus. These findings seem to support a saliency account of coreference resolution, according to which mentions that are more accessible in memory are retrieved faster. However, as both studies do not control the part of speech of the mentions they study, it could be the case that the found effect is in fact caused by the use of anaphoric pronouns. Indeed, anaphoric pronouns are, by definition, discourse old and therefore their use would be correlated with factors such as mention count, used by Jaffe *et al*. (2018), or the discourse old/new distinction, used by Dubey *et al*. (2013). Nevertheless, it should be noted that even if the effect is provoked by anaphoric pronouns, it does not necessarily mean that it goes against the saliency account.

[3] Using a Bayesian framework has the advantage of fewer convergence problems and less need for statistical corrections.

Indeed, Ariel (1988) for example, claims that anaphoric pronouns are typically used to refer to highly salient discourse referents.

Note that we cannot easily compare Seminck's 2018 results to the ones of Dubey *et al.* (2013) and Jaffe *et al*. (2018). The first reason is that Seminck (2018) only studied anaphorical pronouns, whereas the others studied all types of coreference, without making a distinction between different forms of coreference. A second reason is that the cost metrics are conceptually different. Both Dubey *et al*. (2013) and Jaffe *et al.* (2018) have a simplistic approximation of the coreference relation by taking only the number of mentions of the same entity into account. Seminck (2018) uses a state of the art coreference resolution system, that --- thanks to a neural network architecture --- has learnt coreference on a training corpus, but of which the exact way of decision-making remains unknown. Therefore, the fact that she finds an influence of pronoun resolution on human reading behaviour, shows that metrics based on Information Theory can be relevant for modelling discourse phenomena and also how discourse phenomena influences reading behaviour. The entropy-metric illustrates that the quantity of information plays an important role in the process of pronoun resolution, just like it does in other linguistic processes, such as syntactic parsing, or lexical retrieval, for which surprisal has been shown to be a relevant measure (Mitchell *et al*., 2010). However, Seminck's (2018) model does not give insights into which exact features in the anaphoric relation contribute to more or less cognitive load. The coreference resolution system operates more or less as a black box, in contrast to the studies of Dubey *et al*. (2013) and Jaffe *et al*. (2018).

The studies we discussed in this article are very preliminary, but there are already some important lessons we can learn from them. First, it is interesting to see that all three studies observed effects directly on the mentions[4] and that the reading measurements were rather early for Jaffe *et al*. (2018) and Seminck (2018) (self-paced reading time and fixation behaviour). It cannot be excluded that later processing measurements do not show any effect, but it is important to notice that pronoun and coreference resolution is a process that starts immediately upon encountering the mention. A second lesson we can learn is explained well in Jaffe *et a*l.'s (2018) study: they describe the effect of mention frequency as small and underline that the effects of pronoun and coreference resolution might be exaggerated in the setting of psycholinguistic experiments because the items are artificial. Indeed, in the studies of Dubey *et al.* (2013) and Seminck (2018) the effects of pronoun and coreference resolution are also small compared to the other factors of the models.

The studies we reviewed show that we can use naturalistic reading corpora to study discourse phenomena such as pronoun and coreference resolution. This enables multiple perspectives. First of all, we can use these corpora to test implementation of dominant theories of coreference resolution, such as Accessibility Theory (Ariel, 1988), Centering Theory (Grosz *et al*., 1995) and the more recently developed Bayesian Model of Pronoun Resolution (Kehler & Rohde, 2013, 2018). Indeed, evidence for these theories has been gathered during psycholinguistic experiments and some corpus studies, but --- except for the studies we reviewed in this article --- we are not aware of other corpus studies that included physiological measurements such as reading times. Testing these theories on naturalistic

[4] Jaffe *et al*. (2018) also found effects one word after the mention.

reading corpora could, on the one hand, provide insight into the robustness of these theories and, on the other hand, help to develop these theories further.

A second important perspective is the development of broad coverage models of language processing, namely, models that take into account all levels of linguistic processing and simulate human performance as closely as possible. In a way, Dubey *et al*. (2013) made a significant beginning with their research project, combining syntactic surprisal and coreference surprisal in one measurement. But we believe that there could be even more linguistic levels included, such as semantics. We also believe that the discourse presentation they used is still very simplistic and could be enhanced. Broad coverage models developed on naturalistic reading corpora could be a valuable contribution to theories of linguistic processing.

§4 Conclusion

In this article we reviewed three studies that looked at pronoun and coreference resolution in naturalistic reading corpora. Whereas these corpora are used more and more to study syntactic processing, the number of studies involving discourse phenomena remains low. There is nevertheless a lot of potential in using these corpora for this purpose. In contrast to psycholinguistic studies, the discourse in naturalistic reading corpora is not artificial and of considerable length.

The studies we reviewed all demonstrate, on naturalistic reading corpora, that pronoun and coreference resolution influences reading behaviour. We see this result as encouragement to develop further models that incorporate discourse phenomena that are robust and of broad coverage. Moreover, we posit that the results can help, in the future, to develop the implementation of well-known theories, such as Centering Theory and Accessibility Theory, that can be evaluated on naturalistic reading corpora.

References:

Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics*, *24*(01), 65–87.

Armeni, K., Willems, R. M., & Frank, S. L. (2017). Probabilistic language models in cognitive neuroscience: Promises and pitfalls. *Neuroscience & Biobehavioral Reviews*, *83*, 579–588.

Barrett, M., & Søgaard, A. (2015). Reading behavior predicts syntactic categories. *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, 345–349.

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157*, 81–94.

Brysbaert, M., & Vitu, F. (1998). Word skipping: Implications for theories of eye movement control in reading. In *Eye guidance in reading and scene perception* (pp. 125–147). Elsevier.

Clark, H. H., & Sengul, C. (1979). In search of referents for nouns and pronouns. *Memory & Cognition*, *7*(1), 35–41.

Crawley, R. A., Stevenson, R. J., & Kleinman, D. (1990). The use of heuristic strategies in the

interpretation of pronouns. *Journal of Psycholinguistic Research*, *19*(4), 245–264.
Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.
Downey, A. (2013). *Think Bayes: Bayesian statistics in Python*. O'Reilly Media, Inc.
Dubey, A., Keller, F., & Sturt, P. (2013). Probabilistic modeling of discourse-aware sentence processing. *Topics in Cognitive Science*, *5*(3), 425–451.
Ehrlich, K., & Rayner, K. (1983). Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior*, *22*(1), 75–87.
Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*(6), 829–834.
Frank, S. L., Koppen, M., Noordman, L. G., & Vonk, W. (2007). Coherence-driven resolution of referential ambiguity: A computational model. *Memory & Cognition*, *35*(6), 1307–1322.
Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S., & Fedorenko, E. (2018). The Natural Stories Corpus. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 76-82.
Gernsbacher, M. A. (1990). *Language Comprehension as Structure Building*. Hillsdale.
Gernsbacher, M. A., & Hargreaves, D. J. (1988). Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language*, *27*(6), 699–717.
Gernsbacher, M. A., Hargreaves, D. J., & Beeman, M. (1989). Building and accessing clausal representations: The advantage of first mention versus the advantage of clause recency. *Journal of Memory and Language*, *28*(6), 735–755.
Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, Language, Brain*, 95–126.
Grosz, B. J., Joshi, A. K., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. *Proceedings of the 21st Annual Meeting on Association for Computational Linguistics*, 44–50.
Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, *21*(2), 203–225.
Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, 1–8.
Jaffe, E., Shain, C., & Schuler, W. (2018). Coreference and focus in reading times. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, 1–9.
Järvikivi, J., van Gompel, R. P., Hyönä, J., & Bertram, R. (2005). Ambiguous pronoun resolution contrasting the first-mention and subject-preference accounts. *Psychological Science*, *16*(4), 260–264.
Kehler, A., & Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, *39*(1–2), 1–37.
Kehler, A., & Rohde, H. (2018). Prominence and coherence in a Bayesian theory of pronoun interpretation. *Journal of Pragmatics*.
Kennedy, A., Hill, R., & Pynte, J. (2003). The Dundee Corpus. *Proceedings of the 12th European Conference on Eye Movement*.
Koornneef, W. A. (2008). *Eye-catching Anaphora* [PhD Thesis]. Universiteit Utrecht.
Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, *41*(5), 1202–1241.
Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end neural coreference resolution.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 188–197.
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.
Maratsos, M. P. (1973). The effects of stress on the understanding of pronominal co-reference in children. *Journal of Psycholinguistic Research*, *2*(1), 1–8.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: WH Freeman.
Mitchell, J., Lapata, M., Demberg, V., & Keller, F. (2010). Syntactic and semantic factors in processing difficulty: An integrated measure. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 196–206.
R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org
Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372.
Seminck, O. (2018). *Cognitive Computational Models of Pronoun Resolution* [PhD Thesis]. Sorbonne Paris Cité, Université Paris Diderot (Paris 7).
Seminck, O., & Amsili, P. (2018). A gold anaphora annotation layer on an eye movement corpus. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 3518–3522.
Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.
Sheldon, A. (1974). The role of parallel function in the acquisition of relative clauses in English. *Journal of Verbal Learning and Verbal Behavior*, *13*(3), 272–281.
Smyth, R. (1994). Grammatical determinants of ambiguous pronoun resolution. *Journal of Psycholinguistic Research*, *23*(3), 197–229.
Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, *48*(3), 542–562.
Thomas, J. A., & Cover, T. (2006). *Elements of Information Theory* (2nd ed.). Wiley New York.
Van Deemter, K., & Kibble, R. (2000). On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, *26*(4), 629–637.
Van Gompel, R. P., & Majid, A. (2004). Antecedent frequency effects during the processing of pronouns. *Cognition*, *90*(3), 255–264.
Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2015). Prediction during natural language comprehension. *Cerebral Cortex*, *26*(6), 2506–2516.